

DOCKET NO. 02-C-129

PATENT

METHOD AND SYSTEM FOR PROVIDING CASCADED TRIE-BASED NETWORK
PACKET SEARCH ENGINES

Inventor(s) :

Nicholas J. Richardson
740 Archer Street
San Diego
San Diego County
California 92109
Citizen of the United Kingdom

Assignee:

STMicroelectronics, Inc.
1310 Electronics Drive
Carrollton, Texas 75006-5039

William A. Munck
DAVIS MUNCK, P.C.
900 Three Galleria Tower
13155 Noel Road
Dallas, Texas 75240
(972) 628-3600

**METHOD AND SYSTEM FOR PROVIDING CASCADED TRIE-BASED NETWORK
PACKET SEARCH ENGINES**

TECHNICAL FIELD OF THE INVENTION

5 [0001] The present invention is directed in general to network routers and, more specifically, to a method and system for providing cascaded trie-based network packet search engines.

10 **BACKGROUND OF THE INVENTION**

[0002] Network routers for packet-based communications protocols such as Internet Protocol (IP) direct incoming information to the next neighbor along a route to the intended destination for the packet. To do this, typically each router
15 along the route performs route address prefix (normally referred to as just "prefix") look-up operations on a prefix (or routing) table to determine the appropriate next hop address for the destination IP prefix. Such operations are generally performed by either an embedded network processor or, more commonly, a
20 separate network search engine.

[0003] Originally the hardware for network search engines employed content addressable memory (CAM), a type of memory consisting of a bit comparator and two memory elements, one

storing data and the other storing a compare mask. The CAM compares incoming data with the value stored in the data memory under the control of the mask value, which may be programmed to override the comparison result to "always match". In operation, a CAM-based network search engine functions by storing all prefixes of a routing table in a CAM array in a specific, prioritized order, with each prefix's associated next hop information stored in a corresponding location in another memory. During prefix look-up, a key is placed on the comparand (compare operand) bus of the CAM array and compared against all prefixes in the memory. The array of match results from all comparisons is sent through a priority logic unit to determine the highest priority match, with the winning match used to address the next hop memory from which the corresponding next hop information is read and returned.

[0004] More recently, software-based network search engines employing a general-purpose processor and a normal memory have been developed. Within such devices, the processor performs prefix searches with a series of memory read and comparison operations. The routing table prefixes and next hop information are typically stored in the memory in data structures built according to one of various software algorithms developed to

reduce memory usage in storing the routing table and the number of memory accesses during look-up. For these purposes, a multi-bit trie and the corresponding algorithm are among the data structures and algorithms that achieve the best data compression with a bounded number of memory accesses for search operations.

[0005] A trie is a tree-based data structure built to represent binary strings, where each bit or group of bits in the string determines the direction taken among branches within the tree. A binary trie proceeds bit-by-bit and has at most two branches from each node, while a multi-bit trie consumes multiple bits at a time and has several branches at each node, each branch leading to the next level. The number of bits consumed or examined during branch selection at each node is referred to as a stride. A uniform width stride trie is a trie with all strides having the same width, except possibly the last stride, which may be the remainder of the prefix length after being divided by the stride width.

[0006] Generally, the multi-bit trie algorithm works by storing and retrieving prefixes in a uniform stride width trie, grouping all branches in the same level with the same parent stride value into a table, referred to as a trie table. At each level, the corresponding stride value provides an index into a

trie table entry containing the information needed to get to the next level. A multi-bit trie has the advantage that prefixes with common high order bits (strides) will share the same parent trie tables, reducing the memory required to store the prefixes.

5 [0007] Routing table look-up is also performed in same width strides, with the value of the next level stride from the input search key (typically an IP address of 32 or 64 bits) decoded and processed together with the associated data field in the stride value's parent table entry. If a stored route with the
10 same prefix stride value is determined to exist within the trie, an index is calculated using the information in the parent table, then the search continues using the table pointer and the calculated index to form an address leading to the next level trie table entry. If a match is not found, the search
15 terminates without success. If a search reaches an end node and a match is found, the search is successful and the associated next hop information is read from the next hop table.

[0008] In conventional processing systems that operate as described above, multiple stored prefixes may match a single
20 destination address that is used as the search key during a look-up operation. This is due to the fact that any n-bit prefix is defined to match the search key even if the search key

is longer than the prefix, provided the complete n-bit prefix matches the first n bits of the search key. Thus, for example, a 2-bit prefix and a 16-bit prefix may both match a 32-bit search key if the 2-bit prefix matches the first two bits of the
5 search key and the 16-bit prefix matches the first sixteen bits of the search key. In order to resolve multiple matches, therefore, the search engine selects only the data associated with the longest matching prefix when responding to the look-up operation.

10 [0009] However, selecting between multiple matches in this way fails to address the problem of matches found within multiple search engines. Continuing the above example, if the 2-bit prefix was in a first search engine and the 16-bit prefix was in a second search engine, both search engines would respond
15 to the look-up operation, creating a conflict.

[0010] Current processing systems with multiple search engines solve this problem by requiring the network processor to perform separate searches within each search engine and then to determine which of the multiple responses to accept. However,
20 this solution is unsatisfactory as it results in a significant increase in look-up time.

SUMMARY OF THE INVENTION

[0011] In accordance with the present invention, a method and system for providing cascaded trie-based network packet search engines are provided that substantially eliminate or reduce disadvantages and problems associated with conventional methods and systems.

[0012] According to one embodiment of the present invention, a method for providing cascaded trie-based network packet search engines is provided. The method includes receiving a search command at one of the network packet search engines. The search command comprises a specific search key. A determination of a longest prefix match based on the specific search key is made at the network packet search engine. A determination is made at the network packet search engine regarding whether the longest prefix match comprises an overall longest prefix match among the cascaded network packet search engines such that any of the cascaded network packet search engines may comprise the overall longest matching prefix independently of position relative to the other cascaded network packet search engines.

[0013] According to another embodiment of the present invention, a method for operating a network processing unit coupled to a plurality of network packet search engines is

provided. The method includes sending a search command to the plurality of network packet search engines and receiving a response to the search command from a single one of the network packet search engines. The single network packet search engine
5 comprises an overall longest prefix match among the plurality of network packet search engines independently of position relative to the other network packet search engines.

[0014] According to yet another embodiment of the present invention, a network packet search engine coupled to at least
10 one other network packet search engine is provided. The network packet search engine is operable to receive a search command. The search command comprises a specific search key. The network packet search engine is also operable to determine a longest prefix match based on the specific search key, to determine
15 whether the longest prefix match comprises an overall longest prefix match among the plurality of network packet search engines, and to respond to the search command based on the determination that the network packet search engine comprises the overall longest prefix match.

20 [0015] Technical advantages of one or more embodiments of the present invention include providing an improved method for providing cascaded trie-based network packet search engines. In

a particular embodiment, a longest prefix match among multiple network packet search engines may be determined by the network packet search engines themselves. Accordingly, only the network packet search engine with the longest prefix match responds to
5 look-up operation and storage requests from the network processing unit. As a result, the network processing unit does not have to resolve conflicts created by matches found within multiple network packet search engines, which decreases look-up time. In addition, the method allows for the maximization of
10 memory utilization in response to storage requests from the network processing unit.

[0016] Other technical advantages will be readily apparent to one skilled in the art from the following figures, description, and claims.

15 [0017] Before undertaking the DETAILED DESCRIPTION OF THE INVENTION, it may be advantageous to set forth definitions of certain words and phrases used throughout this patent document: the terms "include" and "comprise," as well as derivatives thereof, mean inclusion without limitation; the term "or," is
20 inclusive, meaning and/or; the phrases "associated with" and "associated therewith," as well as derivatives thereof, may mean to include, be included within, interconnect with, contain, be

contained within, connect to or with, couple to or with, be communicable with, cooperate with, interleave, juxtapose, be proximate to, be bound to or with, have, have a property of, or the like; and the term "controller" means any device, system or
5 part thereof that controls at least one operation, such a device may be implemented in hardware, firmware or software, or some combination of at least two of the same. It should be noted that the functionality associated with any particular controller may be centralized or distributed, whether locally or remotely.

10 Definitions for certain words and phrases are provided throughout this patent document, those of ordinary skill in the art should understand that in many, if not most instances, such definitions apply to prior, as well as future uses of such defined words and phrases.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] For a more complete understanding of the present invention and its advantages, reference is now made to the following description taken in conjunction with the accompanying drawings, wherein like reference numerals represent like parts, in which:

[0019] FIGURE 1 is a block diagram illustrating a processing system comprising cascaded trie-based network packet search engines in accordance with one embodiment of the present invention;

[0020] FIGURE 2 is a block diagram illustrating one of the network packet search engines of FIGURE 1 in accordance with one embodiment of the present invention; and

[0021] FIGURE 3 is a flow diagram illustrating a method for providing the cascaded trie-based network packet search engines of FIGURE 1 in accordance with one embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[0022] FIGURES 1 through 3, discussed below, and the various embodiments used to describe the principles of the present invention in this patent document are by way of illustration only and should not be construed in any way to limit the scope of the invention. Those skilled in the art will understand that the principles of the present invention may be implemented in any suitably arranged processing system.

[0023] FIGURE 1 is a block diagram illustrating a processing system 100 comprising at least two cascaded trie-based network packet search engines (NPSEs) 102 in accordance with one embodiment of the present invention. In addition to the network packet search engines 102, the processing system 100 comprises a system controller or network processing unit (NPU) 104, a backplane 106, a network interface 108, and external memories 110. According to one embodiment, the processing system 100 may implement a portion of an Internet Protocol (IP) network router.

[0024] The network processing unit 104 is coupled to the network packet search engines 102, which are each coupled to a corresponding one of the external memories 110. As used herein, "each" means every one of at least a subset of the identified items. The network processing unit 104 is also coupled to the

backplane 106 and to a network through the network interface 108.

[0025] Each network packet search engine 102 is operable to receive addresses, commands and/or prefixes for the next hop address from the network processing unit 104 and to return the next hop address to the network processing unit 104. In order to perform these tasks, each network packet search engine 102 is operable to transmit data table memory addresses, together with a data pointer to the next hop address within a data table, to the corresponding external memory 110. In addition, each network packet search engine 102 is operable to receive the next hop address that is identified by the address and data pointer from the corresponding external memory 110.

[0026] As described in more detail below in connection with FIGURES 2 and 3, each network packet search engine 102 is operable to receive input data from and transmit output data to any adjacent network packet search engines 102 in order to determine the longest prefix match (LPM) among multiple network packet search engines 102. By communicating with the other network packet search engines 102, each of the network packet search engines 102 is operable to compare its longest prefix match with other matches that may have been found by other

network packet search engines 102. The network packet search engine 102 that determines that it has found the longest prefix match may then respond to the network processing unit 104.

[0027] The network packet search engines 102 may comprise
5 either peripheral network packet search engines 102a or central network packet search engines 102b. A peripheral network packet search engine 102a is coupled to only one adjacent network packet search engine 102, while a central network packet search engine 102b is coupled to two network packet search engines 102.

10 [0028] Although the illustrated embodiment comprises four network packet search engines 102, it will be understood that the processing system 100 may comprise any suitable number of network packet search engines 102 without departing from the scope of the present invention. Thus, for embodiments of the
15 processing system 100 that comprise more than two network packet search engines 102, the processing system 100 comprises two peripheral network packet search engines 102a and the remaining network packet search engines 102 are central network packet search engines 102b. For embodiments of the processing system
20 100 that comprise only two network packet search engines 102, those network packet search engines 102 are peripheral network packet search engines 102a and the processing system 100

comprises no central network packet search engines 102b.

[0029] According to one embodiment, the network packet search engines 102 may be coupled to the network processing unit 104 by a high-speed 36-bit network processor unit interface and to the external memories 110 by high-speed 36-bit static random access memory (SRAM) interfaces. For this embodiment, each interface may be implemented by a quad data rate (QDR) architecture, with the network processing unit 104 as the master for the NPU interface and the external memories 110 as the slaves for the SRAM interfaces.

[0030] FIGURE 2 is a block diagram illustrating one of the network packet search engines 102 in accordance with one embodiment of the present invention. The network packet search engine 102 is operable to receive input data from and transmit output data to any adjacent network packet search engines 102.

[0031] The network packet search engine 102 comprises at least one prefix table 200 that is operable to store prefixes for the network packet search engine 102. Each prefix stored in a prefix table 200 is operable to identify an address stored within the corresponding external memory 110.

[0032] The network packet search engine 102 also comprises a longest match right input (LMRI) node 202 for receiving the

length of the longest prefix match found by the network packet search engines 102 to the right of the network packet search engine 102 and a longest match left input (LMLI) node 204 for receiving the length of the longest prefix match found by the network packet search engines 102 to the left of the network packet search engine 102.

[0033] The network packet search engine 102 also comprises a longest match right output (LMRO) node 206 for transmitting the length of the longest prefix match found by the network packet search engines 102 to the right of the network packet search engine 102 or by the network packet search engine 102, whichever is longer, and a longest match left output (LMLO) node 208 for transmitting the length of the longest prefix match found by the network packet search engines 102 to the left of the network packet search engine 102 or by the network packet search engine 102, whichever is longer.

[0034] The network packet search engine 102 also comprises a match valid node for each of the input and output nodes to indicate whether or not the data at the corresponding node is valid. Thus, the network packet search engine 102 comprises a match valid right input (MVRI) node 212 that corresponds to the LMRI node 202, a match valid left input (MVLI) node 214 that

corresponds to the LMLI node 204, a match valid right output (MVRO) node 216 that corresponds to the LMRO node 206, and a match valid left output (MVLO) node 218 that corresponds to the LMLO node 208.

5 [0035] FIGURE 3 is a flow diagram illustrating a method for providing cascaded trie-based network packet search engines 102 in accordance with one embodiment of the present invention. The method begins at step 300 where the network processing unit 104 sends a search command to each of the network packet search
10 engines 102.

[0036] The search command comprises a retrieval command, an insertion command and/or any other suitable command which initiates a process to find a longest prefix match. Each search command comprises a specific search key. The retrieval command
15 comprises a command to initiate a search for a longest prefix match within the network packet search engines 102 that may be used, for example, to identify a next hop address within one of the external memories 110 associated with the specific search key. The insertion command comprises a command to initiate a
20 search for a longest prefix match within the network packet search engines 102 in order to select one of the network packet search engines 102 for storing the specific search key and its

corresponding external memory 110 for storing a new address associated with the specific search key.

[0037] For the following steps (steps 302-308), it will be understood that these steps are performed by each of the network
5 packet search engines 102 independently.

[0038] At step 302, a network packet search engines 102 searches its prefix table 200 for prefixes matching the specific search key. At decisional step 304, the network packet search engine 102 makes a determination regarding whether or not any
10 matches were found. If no matches were found, the method follows the No branch from decisional step 304 to step 306. At step 306, the length of the longest prefix match is set to zero for that network packet search engine 102.

[0039] Returning to decisional step 304, if one or more
15 matches were found, the method follows the Yes branch from decisional step 304 to step 308. At step 308, that network packet search engine 102 determines the length of its longest prefix match.

[0040] At step 310, each peripheral network packet search
20 engine 102a transmits the length of its longest prefix match to its adjacent network packet search engine 102, which may be another peripheral network packet search engine 102a or a

central network packet search engine 102b depending on the number of network packet search engines 102 in the processing system 100.

[0041] For the following steps (steps 312-328), it will be understood that these steps are performed by each of the network packet search engines 102 independently.

[0042] At step 312, a network packet search engine 102 receives longest match input data at one of the longest match input nodes (the LMRI node 202 or the LMLI node 204). The longest match input data comprises the length of the longest prefix match for one of the other network packet search engines 102, as described above in connection with FIGURE 2. The network packet search engine 102 also receives match valid input data at the corresponding match valid input node (the MVRI node 212 or the MVLI node 214) that indicates that the longest match input data is valid.

[0043] At step 314, the network packet search engine 102 compares the longest match input data received at the longest match input node 202 or 204 to the length of its own longest prefix match. At decisional step 316, the network packet search engine 102 makes a determination regarding whether or not the longest match input data is greater than the length of its own

longest prefix match. If the network packet search engine 102 determines that the longest match input data is greater than the length of its own longest prefix match, the method follows the Yes branch from decisional step 316 to step 318.

5 [0044] At step 318, the network packet search engine 102 transmits the longest match input data to an adjacent network packet search engine 102. This data is transmitted as longest match output data at the longest match output node 206 or 208 corresponding to the longest match input node 202 or 204 at
10 which the longest match input data was received. Thus, if the longest match input data was received at the LMRI node 202, the data is transmitted as longest match output data at the LMRO node 206. Similarly, if the longest match input data was received at the LMLI node 204, the data is transmitted as
15 longest match output data at the LMLO node 208.

[0045] In addition, the network packet search engine 102 also transmits match valid output data at the corresponding match valid output node (the MVRO node 216 or the MVLO node 218) that indicates that the longest match output data being transmitted
20 is valid.

[0046] Returning to decisional step 316, if the network packet search engine 102 determines that the longest match input

data is not greater than the length of its own longest prefix match, the method follows the No branch from decisional step 316 to step 320.

[0047] At step 320, the network packet search engine 102 transmits the length of its own longest prefix match to an adjacent network packet search engine 102. This data is transmitted as longest match output data at the longest match output node 206 or 208 corresponding to the longest match input node 202 or 204 at which the longest match input data was received. Thus, if the longest match input data was received at the LMRI node 202, the data is transmitted as longest match output data at the LMRO node 206. Similarly, if the longest match input data was received at the LMLI node 204, the data is transmitted as longest match output data at the LMLO node 208.

[0048] In addition, the network packet search engine 102 also transmits match valid output data at the corresponding match valid output node (the MVRO node 216 or the MVLO node 218) that indicates that the longest match output data being transmitted is valid.

[0049] At decisional step 322, the network packet search engine 102 makes a determination regarding whether or not all expected longest match input data has been received. Thus, the

peripheral network packet search engines 102a determine whether or not longest match input data has been received from one adjacent network packet search engine 102 and the central network packet search engines 102b, if any, determine whether or
5 not longest match input data has been received from two adjacent network packet search engines 102.

[0050] If all longest match input data has not been received, the method follows the No branch from decisional step 322 and returns to step 312 where the network packet search engine 102
10 receives longest match input data at one of the longest match input nodes. However, if all longest match input data has been received, the method follows the Yes branch from decisional step 322 to decisional step 324.

[0051] At decisional step 324, the network packet search
15 engine 102 makes a determination regarding whether or not the length of its longest prefix match is greater than the length of any LPMs found by other network packet search engines 102. Thus, if the network packet search engine 102 determines that its longest prefix match is longer than all the longest match
20 input data received from other network packet search engines 102, that network packet search engine 102 comprises the overall longest prefix match among the network packet search engines 102

and the method follows the Yes branch from decisional step 324 to step 326. In this way, any of the network packet search engines 102 may comprise the overall longest matching prefix independently of its position relative to the other network packet search engines 102.

[0052] At step 326, the network packet search engine 102 responds to the search command received from the network processing unit 104, at which point the method comes to an end. According to one embodiment, the network packet search engine 102 actively responds to the network processing unit 104. For this embodiment, if the search command comprises a retrieval command, the network packet search engine 102 uses its longest prefix match to retrieve data from the corresponding external memory 110 and transmits that data to the network processing unit 104. Similarly, if the search command comprises an insertion command, the network packet search engine 102 stores data from the network processing unit 104 in its corresponding external memory 110, in addition to storing the specific search key in its own prefix table 200.

[0053] According to another embodiment, the network packet search engine 102 passively responds to the network processing unit 104. For this embodiment, any one of the network packet

search engines 102 may provide a "not ready" signal to the network processing unit 104 until one of the network packet search engines 102 determines that its longest prefix match is the overall longest prefix match in decisional step 324. Then, at step 326, that network packet search engine 102, which has stored its result (either data retrieved from the corresponding external memory 110 for a retrieval command or identifying itself as the location for storing a new prefix for an insertion command) in its output buffer, provides a "ready" signal to the network processing unit 104. Thus, for this embodiment, the network processing unit 104 may retrieve the data from the output buffer instead of having the data actively transmitted from the network packet search engine 102.

[0054] For this passive embodiment, the network packet search engine 102 that provides the "not ready" signal and the network packet search engine 102 that provides the "ready" signal, along with the appropriate data in its output buffer, are synchronized such that the hand-off from one to the other occurs during a clock cycle. This prevents the possibility of having either no network packet search engines 102 or multiple network packet search engines 102 trying to communicate with the network processing unit 104 at any one time.

[0055] Returning to decisional step 324, if the network packet search engine 102 determines that its longest prefix match is not longer than all the longest match input data received from other network packet search engines 102, the
5 method follows the No branch from decisional step 324 to step 328. At step 328, the network packet search engine 102 may store a mock result in its output buffer as a position holder in order to ensure that correct result ordering is maintained, and the method comes to an end.

10 [0056] Although the present invention has been described with several embodiments, various changes and modifications may be suggested to one skilled in the art. It is intended that the present invention encompass such changes and modifications as fall within the scope of the appended claims.